Advanced Astrophysical Algorithms to Novel Supercomputing Hardware

PI: Robert Brunner, U of Illinois

Astronomy has become a data-rich field, and the future promises to only increase the data volume. NASA has a long, and distinguished track record of funding scientifically important missions that are in large part contributing to the quantity and quality of data being analyzed to gain fundamental insights into everything from our own planet to the origins of our universe. Our ability as scientists, however, to keep pace with this data flood is not sufficient. In response, many researchers, some in proposals to this AISR program, have developed innovative, cutting-edge statistical and computer science techniques and and tools to improve our ability to analyses the data in hand. Even these efforts have not been adequate; we somehow must find mechanisms for improving the performance and robustness of our applications; otherwise we will not only fail to capitalize on the richness of the data we currently have in hand, but we we also will fall further behind with the imminent arrival of petabyte datsets. This problem is not tied to one particular branch of NASAï¿½s operating strategy, but crosses all scientific and engineering disciplines. In this proposal, we outline an interdisciplinary collaboration that we feel has the potential to revolutionize the way in which complex algorithms are applied to astrophysical datasets. Our collaboration of domain specialists have identified a means by which the analysis of large datasets can be increased by a factor of 100 or more over that provided by the traditional use of supercomputers alone. We will combine the expertise of the Laboratory for Cosmological Data Mining, the NCSA Innovative Systems Laboratory (ISL) and the NCSA Automated Learning Group (ALG) to analyze terascale datasets, such as that from the Sloan Digital Sky Survey (SDSS). As an initial proof-of-concept, we will produce the largest and most accurately classified catalog of objects to date, using instance-based learning, and, subsequently use this classified catalog to calculate cosmological statistics such as N-point correlation functions. Both of these data analyses, using the most accurate statistical approaches are otherwise intractable. The ISL has available cutting-edge hardware in the form form of high performance reconfigurable systems which enable the speedup described. Crucially, given the experimental nature of current non-embedded applications which this technology, we also have the expertise and proven results to make this possible. Tests have shown that instance-based learning, in particular when combined with the results from other algorithms, results in superior object classification, classification, but the method remains relatively unexplored in astronomy due to its intractability. The PI and two Co-Is have joint appointments between the Department of Astronomy and NCSA at the University of Illinois, Urbana-Champaign, and have previously found that technological advances often benefit from the proximity of these institutions. The cyberinfrastructure to produce the catalogs is already in place, in the form of the Data-to-Knowledge Toolkit developed by the ALG. Indeed, it has already been used to classify the 141 million objects in the SDSS Data Release 3, a 70 GB dataset, but to to realize the full potential of the available facilities and expertise requires the interdisciplinary collaboration outlined herein. The ISL is keen to gain partners who have real-world applications which may benefit from reconfigurable computing, either in industry or academia. The relevance to NASA and and AISR goals is the advancement of computer hardware and software in the field of high performance computing, cyberinfrastructure and machine learning. The use of FPGAs/reconfigurable computing will make tractable other data intensive problems in astronomy such as Fast Fourier Transforms (e.g., for the analysis of Planck mission data), or object classification and statistical analysis within the anticipated petascale datasets of missions such as the LSST.

Parallel-Processing Astrophysical Image-Analysis Tools

PI: Kenneth Mighell, NOAO

This proposal describes a three-year research plan to develop and implement several parallel-processing astrophysical image-analysis tools for two of NASA's Great Observatories: the Hubble Space Telescope Telescope and the Spitzer Space Telescope. These new data-analysis software tools will be written with the goal of enhancing the scientific return of these missions. This project will improve and extend the AISR-funded enabling image-processing technology of the Principal Investigator's MATPHOT algorithm for precise and accurate stellar photometry and astrometry with discrete Point Spread Functions. The PI has recently demonstrated (in the 11 August 2005 issue of the Monthly Notices of the the Royal Astronomical Society) that the current C implementation of the MATPHOT algorithm can achieve millipixel relative astrometry and millimag photometric precision with complicated space-based space-based discrete Point Spread Functions imaged onto imperfect detectors with large intrapixel quantum efficiency variations. This project combines state-of-the-art astrophysical image processing techniques with the enabling technology of Beowulf clusters which offer excellent cost/performance ratios for computational power. The PI will use the project software to investigate the possibility of significantly improving the current state-of-the-art of space-based optical and infrared stellar photometry and astrometry by analyzing existing archival imaging data from the NASA mission data archives for HST and SST. This project is structured into three one-year phases. The major software products will be posted (at the end of each one-year phase) on web sites dedicated to this project on the web servers of the National Optical Astronomy Observatory.

Bayesian Source Separation for Astrophysical Spectra: Application to PAHs

PI: Kevin Knuth, NASA Ames

We propose to apply Bayesian source separation techniques to a very important subset of astrophysical spectra, the emission spectra of PAHs. The PAH species, ubiquitous in our Galaxy and in external galaxies, are one of the most important molecular species in the pathway to life. Their spectra are the result of the blending of the complex spectra of individual PAH species, possibly at different temperatures, and to date they have defied all attempts to unscramble them into individual contributors. We propose to apply Bayesian source separation for the first time in order to quantitatively determine which of the many hundreds of individual PAH species contribute most significantly to the combined emission spectra of PAHs. Our research will lead to powerful new tools for decomposing, analyzing and characterizing the spectra of these critical species. Furthermore, these tools will be readily generalizable to a wide array of spectral analysis problems ranging across astrophysics, Earth science and biology.

Large Scale on Demand Cross-Matching with Open SkyQuery

PI: Aniruddha Thakar, Johns Hopkins U

Summary This proposal seeks to develop an advanced and versatile all-sky cross-matching engine for large distributed astronomical datasets by bringing together and leveraging the results of two previously funded AISR projects. We wish to extend the present SkyNode and Open SkyQuery functionalities in the following ways: 1. Incorporate parallel data access with Zone partitioning into individual SkyNodes 2. Transport data between SkyNodes and the portal using the Mega Streaming transport protocol developed by Grossman et al. 3. Replace the current synchronous query execution with an asynchronous asynchronous system that is compatible with emerging IVOA standards. We will apply the Zone-based partitioning and parallelism that we developed with prior AISR funding to individual SkyNodes in Open SkyQuery (also developed with AISR funding) so that we can parallelize the data access and the cross-match computation by distributing the data and workload among a cluster of database servers. In order to move the cross-match data quickly between SkyNodes, we will incorporate incorporate the mega-streaming network transport protocol for fast data movement developed at NCDM/UIC. Asynchronous workflow is the third main enhancement that we will make. The size and complexity of the large scale cross-matches will make synchronous response back to the Web client impracticable and inefficient. Asynchronous query execution will require identification of users, tying into the IVOA single sign-on (SSO) protocol. The large outputs resulting from large cross-matches will require direct routing to storage destinations like VOStores. In addition to producing a large-area cross-matching facility that will allow the NVO to truly federate large distributed archives, this project will also provide high-speed data access to all large datasets supported by the NVO and NASA. It will enable unprecedented science that can be used to constrain cosmological n-point correlation function and power spectrum calculations, determine the AGN luminosity function, find galaxy clusters, etc. These studies will increase the science return from current NASA missions and promote the NASA strategic objective of exploring the universe to understand its origin, structure, evolution, and destiny. A facility that allows for extensive correlation with other catalogs will be extremely useful for near-earth near-earth and moving object searches, and for TPF searches related to the NASA strategic objective of searching for Earth-like planets.

BioLingua - An Integrated System for the Discovery how Genes Regulate Cell Behavior

PI: Andrew Pohorille, NASA Ames

In order to improve productivity in astrobiology we propose to develop an intuitive, easy to use programming environment that enables scientists to combine their data, knowledge, computational tools and previous, relevant results to make new biological discoveries. We call this environment ï¿½BioLingua/CDï¿½ (BioLingua for Collaborative Discovery). BioLingua/CD will have a number of fundamental features. It will contain an efficient, easy to use, general purpose programming language augmented with a rich set of biology-specific functionalities, and transparent access to a wide range of basic bioinformatics algorithms, advanced statistical tools and databases. They include genomic, proteomic, metabolic and gene expression data, as well as relevant higher-level knowledge, such as the Gene Ontology, and representations of metabolic and regulatory pathways. All these capabilities, which enable scientists to access and interpret relevant data and knowledge, will be available through a common framework. At the center of BioLingua will be state-of-the art tools for discovering causal relations in gene regulation and metabolic networks and modeling relations between gene expression and metabolism, subject to constraints obtained from previously acquired data and knowledge. The environment will be highly interactive, web-based and equipped with a flexible and efficient visual programming language. The environment will provide tools for collaboration among biologists and a means for sharing results with the community. Using these capabilities the biologist will be able to manipulate and analyze data and knowledge for biological problems with convenience that would be otherwise impossible to achieve. At the end of this project, we will deliver a powerful, easy to use system that will be an essential tool for astrobiologists engaged in genomic, cellular biology and systems systems biology research. This system will provide technology for increasing productivity in all science programs sponsored by the Science Mission Directorate that involve studies of extant or extinct organisms. In particular, it will support missions and research devoted to understanding the evolutionary evolutionary mechanisms and environmental limits of life by determining the molecular, genetic, and biochemical mechanisms that control and limit evolution, metabolic diversity, and the ability of life to survive in space, as defined in Goals 4, 5 and 6 of Astrobiology Roadmap. BioLingua technology has the potential to increase productivity in SMD programs beyond biological domain though its advanced knowledge discovery and data synthesis capabilities, and by fostering interdisciplinary collaborations that span the space science and computer science disciplines.

SkyView: The Visualization Portal for the Virtual Observatory

PI: Thomas McGlynn, NASA Goddard

Over the past decade the SkyView Virtual Telescope has pioneered concepts of common, simple access to multi-wavelength data that have now begun to mature in the Virtual Observatory (VO) efforts underway in many nations. In this proposal we discuss how SkyView can build upon the nascent capabilities and protocols of the VO to provide a whole new way for astronomers to organize, explore and visualize sky data. We propose extending SkyView to unify our view of the heavens in fundamentally new ways. Today, astronomers treat catalogs of objects and images as very different products even though they both are representations of what is in the sky. We propose to enable astronomers to revisualize catalog data helping astronomers to comprehend the huge giga-object catalogs that are now coming on-line in the astronomy community. Today, transformations between representations of large-scale structure, especially for cosmic microwave background, and conventional maps are not easily available to most astronomers and the simple approaches commonly used do not preserve as much information as they might. We propose to build transformation tools for SkyView to efficiently and robustly transform data among all useful astronomical data projections. Today, transforming image geometries is treated as a black art: most users pick the simplest possible algorithms even when they could preserve much more information at little cost in computing power. We propose to to build and distribute a reference library of resampling techniques and algorithms so that users in astronomer and throughout the sciences have easy access to a variety of resampling techniques. As both a consumer and provider of VO datasets, we shall document how other users can effectively build VO-compliant data services. The proposal takes advantage of the confluence of huge new data resources coming on-line in astronomy and the new community standards being promulgated in the Virtual Observatory to enable astronomers to flexibly visualize astronomical data. With our new approaches astronomers can more effectively realize the science potential of NASA missions and projects. Reducing the barriers to understanding and manipulating images and catalogs, our program also enlarges the pool of scientists able to use NASA data for research and makes it possible for the non-science community to participate in NASAï¿½s science programs.

VOClass: Classification and Learning in the Virtual Observatory

PI: Andrew Connolly, U Pittsburgh

Classification and outlier detection are fundamental to all areas of the observational sciences. In astrophysics the interrelations between observed parameters can give insight into the physical processes within the universe. Deviations from these relations can be used to identify new classes of source or astrophysical phenomena. We propose here to develop an automated classification facility for the VO, VOClass, that will integrate state-of-the-art data-mining algorithms with an extensive set of VO-enabled VO-enabled tools that are now available. Classifications will be undertaken utilizing a range of supervised and unsupervised techniques that are designed to scale to the size of current and future surveys. Using VOClass, astronomers will be able to upload images or catalogs, correlate their sources with other multifrequency data sets available through the VO or NASA data archives and return classifications of these sources and identifications of unusual objects based on the measured and correlated properties. The primary goals of VOClass are: (a) to enable fast classification of large, distributed databases, (b) to provide robust statistical descriptions of the interrelation between observables that scale to the size and dimensionality of current and future data sets, (c) to integrate classification and visualization tools available through the Virtual Observatory, (d) to provide the ability ability to adaptively classify sources by learning from user defined inputs and thereby develop new classification schemes, (e) to interface these knowledge discovery tools with the Virtual Observatory through the application of Web Services which will provide a modular access enabling scientists to build build other classifiers that are optimized to a particular task. VOClass will provide a comprehensive facility for source extraction, cross-matching, and classification. Combining classification with state-of-the-art VO tools it will provide an environment for distributed analysis and knowledge discovery that is unprecedented in astronomy. It has, however, many other applications both within astrophysics and in the context of the broader NASA mission. Development of VOClass will directly address the AISR program goal of providing "advances in information science and technology to increase life cycle effectiveness and efficiency of the Science Mission Directorate (SMD)". Beyond astrophysics the techniques we propose to deliver have applications to telemetry data streams, quick look observations and verification of data from satellite images.

HYDRA: A New Paradigm for Astrophysical Modeling, Simulation, and Analysis (contd)

PI: John C Houck, MIT

Regardless of the source or mission, the analysis of astrophysical data invariably relies on the interplay between predictive physical models, a detailed understanding of the observing instruments, and the discriminating comparison of predictions and actual observations. With HYDRA, we are developing a new platform which will provide scientists a more flexible and extensible way of constructing models for astrophysical sources that include geometric information, physical emission and absorption mechanisms, transport processes, and projection effects. We will also provide an interface through which users may link existing astrophysical models to HYDRA. Similarly, modules describing the performance of the instrumentation, be it ground-based telescope or orbiting satellite, will be definable in a mission-independent way to allow realizations of these models in the form of simulated observations. By combining source and instrumentation models, HYDRA can serve as a tool for observational planning, instrument design, and calibration. Simultaneously, HYDRA will provide an analysis environment that allows source models to be compared to existing observations and iteratively adjusted. As part of its design, the HYDRA system will also include advanced visualization capabilities that will provide users with additional diagnostic ability as well as a potentially powerful educational tool.

Scalable Algorithms for Analysis of Megapixel CMB Maps and Large Databases

PI: Istvan Szapudi, U of Hawaii

We propose to develop a set of algorithms for spatial statistical analyses of large astronomical data bases, such a CMB maps, galaxy catalogs, or any point source catalog. We will build on the success of our previous AISR, and tackle computational challenges such as i) fast and accurate estimation of temperature, polarization and lensing correlation functions and power spectra, ii) fast estimation of higher order correlation functions from large CMB maps, angular, redshift and weak lensing surveys. iii) fast estimation of covariance matrices using novel Monte Carlo techniques. In our unique interdisciplinary approach for algorithmic development, we reformulate the statistical problems arising in space astronomy with special attention to computational needs. Our philosophy naturally blends the principles of astronomy, statistics, computer and computational science. Special attention will be payed to the user interface and quality control to ensure wide practical usage. The resulting software package will be useful for accurate analysis of MAP and Planck, or any subsequent megapixel surveys, galaxy surveys, background light correlations, correlations in maps produced in other wavelengths, such as infrared background, as well as cross correlations of all the above.

Enabling Bayesian Inference for the Astronomy Masses

PI: Martin Weinberg, U Massachusetts

The wealth of data being acquired by NASA missions promises detailed tests of scientific theories. Bayesian analysis has the power to efficiently combine information from these diverse sources, rigorously establish confidence bounds on theoretical models, and provide powerful probability-based methods for model selection and complex hypothesis testing. This statistical approach is superior to those commonly used in astronomy. However, it is not currently favored owing to its computational difficulty. To remedy this deficiency, a multi-disciplinary investigator team from the Departments of Astronomy and Computer Science at UMass recently developed the Bayesian Inference Engine (BIE). BIE's success is aided by advances in Markov Chain Monte Carlo (MCMC) algorithms and the availability of commodity parallel hardware. Most importantly, the system frees the scientist from the database mechanics of the ongoing statistical investigation without sacrificing rigor, while providing high-level interaction with the inquiry. To promote wider application and use, we propose key enhancements and stand-alone applications to allow the astronomical community to take immediate advantage of these statistical inference tools. First, we will implement a persistence system. This allows allows the entire inference to be stored in a research log to be later recalled and reused. A persistence system enables what-if exploration, checkpointing, and collaborative investigations. We will enhance the BIE with advanced MCMC algorithms, non-parametric prior distributions and new graphical tools. Second, we will develop three stand-alone "killer" applications that will demonstrate the power of Bayesian inference as well as provide an introduction to using BIE and Bayesian methods for more specialized problems. The applications are: 1) Analysis of Structure in Catalog Databases--a general set of star and galaxy counting routines with I/O interfaces to SQL relational databases (e.g 2MASS and SDSS); 2) Determining Galaxy Components from Images--a galaxy classification tool based on the popular GIM2D package with a BIE back-end that will allow sophisticated inference over image ensembles; and 3) A Rigorous Statistical Basis for Semi-Analytic Models (SAMS)--these widely used multiparameter phenomenological models for galaxy formation make predictions for and use constraints constraints from multiple types of observations. The BIE back-end will enable methodical exploration of the free parameters and allow sophisticated hypothesis testing.

Developing Methods to Incorporate Calibration Uncertainties in Data Analysis

PI: Vinay Kashyap, Harvard/SAO

The aim of this proposal is to develop and make publicly available a set of tools designed to incorporate knowledge of calibration errors into data analysis. We have developed a method to handle in a practical way the effect of uncertainties in instrument response on astrophysical modeling, with specific application to Chandra/ACIS instrument effective area. This groundbreaking work holds great practical promise for a generalized treatment of instrumental uncertainties in not just X-ray spectra but also imaging, or any kind of higher-dimensional analyses. We apply a combination of Principal Component Analysis and Markov chain Monte Carlo methods to calibration uncertainties and include these uncertainties directly into analyses. We propose to develop modular tools that can be easily extended to different instruments and calibration products, and will provide a systematic framework to describe the true errors in the estimated model parameters, even in non-Gaussian regimes. Calibration is the foundation on which all data are analyzed and interpreted. Therefore, any efforts to improve handling of of calibration data and to use it in a better fashion is immediately relevant to Space Science programs. Considerable effort will be saved by reducing the propensity of systematic calibration-based biases from from skewing the results. Our proposed project will also have specific impacts. Firstly, because algorithm development will mostly take place in the context of Chandra data, it will lead directly to an improvement in the quality of the analysis of Chandra data. Secondly, it will result in the useful characterization of a number of calibration products from Chandra, Suzaku, and other current missions. Finally, we will set up groundbreaking standards in the manner and specification of calibration data for future missions.

On-the-fly and Grid Analysis of Astronomical Images for the Virtual Observatory

PI: Andrew Ptak, Johns Hopkins U

We propose to improve upon two existing software packages, XAssist and WESIX. WESIX automatically performs source detection on images that are uploaded by users. XAssist automatically analyzes X-ray astrophysics data, but currently does not have an interface for the Virtual Observatory, and a major goal will be to provide a web service interface to XAssist. The main improvements to XAssist will be added functionality to only analyze a subset of a field (i.e., around the requested position), the capacity to distribute its processing among multiple machines or CPUs, and enhanced speed and reliability. WESIX will be improved to compute calibrated optical fluxes for the detected sources, particularly for Hubble Space Telescope data. The ability to compute upper-limits will be added to both systems, which is particularly important for minimizing bias in statistical studies. We will will also work on the production of a common interface to both systems that will allow the processing of of optical and X-ray data to be requested in the same way, and allow the software to be used in a grid computing environment. This will set a standard for the inclusion of similar systems for other wavelengths, such as radio and infrared data. This project will improve the utilization of NASA archival archival data.

Enabling Massive Scientific Databases Through Automated Schema Design

PI: Jeffrey Gardner, Carnegie Mellon

As massively parallel platforms continue to expand in processor count, simulations can now generate datasets of unprecedented size relative to the processing power of a serial workstation. NASA has pushed this envelope with the Columbia supercomputer at NASA/ARC, which has 10,240 processors with a combined capability exceeding 50 teraflops. With the ability to run on 10 to 100 times as many processors as we could a mere five years ago, one of the greatest challenges scientists face in using these these impressive resources is extracting meaningful scientific knowledge out of the immense datasets that they can generate. Simulations are now being limited in scope not by the capabilities of the computers that run them, but by the scientist's inability to cope with the resultant data flow. The most obvious solution when dealing with massive data volumes is to register it into a database. The difficulty difficulty lies in the amount of computing and access time required to execute database queries: a cosmology simulation that fully utilized Columbia, for example, would generate at least 20 TB of data. The trick to optimizing queries lies in designing an intelligent schema, i.e. a description how the data is actually arranged within the database. Databases of this size mandate optimal schema design. Unfortunately, simulation groups--which are often quite small--rarely have the resources to spend on manually designing efficient database schema. Furthermore, the schema design for simulations is typically much more difficult than for observational datasets: a simulation catalog not only incorporates the time dimension, but it also frequently employs more layers of relationships. Current automated design tools are not useful for our purposes as they rely on aggressively replicating indexes, a strategy that increases the size of the database by factors of 2 to 3. We propose to overcome the barriers to efficient query execution by combining work and ideas of the database design community with the needs of the computational cosmology domain. We will design and implement AutoPart, a system that automatically designs scientific databases by partitioning the tables in the original database according to a representative workload. Compared to conventional index-based techniques, AutoPart removes the need for replication by pre-designing the tables according to the query requirements. AutoPart will maximize performance of scientific queries while eliminating the need for additional space and update overhead, thereby permitting any researcher to interact with their data in the most efficient way possible. possible. Our initial focus will be on enabling the mining of massive cosmological simulations. The results of our work, however, will be applicable far beyond cosmology and will effect any NASA endeavor that can exploit massive datasets.

Statistical Inference for Scientific Instruments: Event Analysis for the Gamma-Ray Large Area Space

PI: Robin Morris, Ames/USRA

Scientific instruments are becoming more complex, are making indirect measurements that require complex interpretation, and are returning enormous quantities of data that require powerful methods of analysis. The Large Area Telescope (LAT) instrument on the Gamma-ray Large Area Space Telescope (GLAST) satellite is a prime example of an instrument where powerful statistical methods are needed at all stages of the data analysis -- the instrument makes indirect measurements of the incident gamma-rays, requiring complex event reconstruction to estimate the directions and energies of the photons; classification is needed to determine whether the event was caused by a gamma-ray or a charged particle; the detection and characterization of celestial sources of gamma-rays requires complex statistical hypothesis testing which is dependent strongly on the statistical characterization of the reconstruction and classification stages. In this proposal we focus on the first of these stages, on which all the others are built. The first task is to estimate the directions and energies of the incoming photons. The tracker element of the LAT is a series of 18 tungsten foils, interleaved with silicon microstrip detectors. Incident gamma-rays interact with the tungsten and are converted into electron-positron pairs. pairs. These charged particles trigger the microstrip detectors. However, the electron and positron are affected by numerous physics processes as they traverse the detector. The primary process that blurs the the response is multiple scattering, which causes the tracks of the electron and positron to deviate from the ideal straight line paths. Other processes result in the ejection of further electrons from the material of the detector, and the production of further gamma-rays by positron annihilation. These gamma-rays can be converted into electron-positron pairs later in the detector. These processes and others produce charged particles that also interact with the silicon microstrips and provoke a response in the detector. The current analysis methodology being developed by the LAT collaboration involves finding the straightest tracks through the microstrip responses, a Kalman filter to estimate the trajectories of the primary electron and positron, and from the trajectories estimating the direction and energy of the photon. We propose an alternative methodology, based on nonparametric estimation using particle filters, that can more effectively use the full physics distributions for multiple scattering and other physics processes, and should give more accurate estimates of the direction and especially the energy of the incident photons and the uncertainties of these estimates - essentially a psf per event, which will enable more powerful and accurate estimation methods to be used in later stages of the data analysis. Improving the estimate of the photon's energy from thetracker response has been identified as an important goal for the event analysis. Preliminary work, funded by the NASA Ames Director's Discretionary Fund has demonstrated the applicability of the methodology to the LAT event analysis problem. This proposal is to further develop the new analysis methodology, and to integrate it into the GLEAM software environment being developed by the GLAST collaboration. A successful outcome would result in a greater science return from the mission deemed of highest priority in its category by the NRC Decadal Review, and also a wider knowledge of a new, powerful statistical methodology amongst the astrophysics and particle physics communities.

Fault-Tolerant Parallel-Processing Astronomical Image-Analysis Tools

PI: Kenneth Mighell, National Optical Astronomy Observatories

This proposal describes a three-year research plan to develop and implement several new fault-tolerant parallel-processing astronomical image-analysis tools suitable for the analysis of stellar imaging data from the current and future NASA astrophysical missions. One such tool will be a fault-tolerant version of the PI's CCD Circular Aperture Photometry (CCDCAP) code for the Hubble Space Telescope's Advanced Camera for Surveys (ACS) instrument. The PI will enhance his existing AISR-funded QLWFPC2 quick-look photometry data-mining code and the MATPHOT precision PSF-fitting CCD stellar photometry and astrometry code by making them fault tolerant as well as faster and more robust. These data-analysis software tools will be written with the goal of enhancing the scientific return of the the HST and the Spitzer Space Telescope missions as well as future planned missions such as the James Webb Space Telescope. Efforts will be made to implement some of the data-mining tools at NASA data archives or mirror sites for HST and SST in order to provide an interactive and more scientifically useful useful user experience while possibly making the serving of archival imaging data more efficient. The fault tolerance of these applications will be achieved by working closely with the LAM/MPI 10.0 team (at the Los Alamos National Laboratory and Indiana University) to implement state-of-the-art fault-tolerant techniques into parallel-processing applications suitable for space-based scientific analysis with a cluster of COTS-based fault-tolerant computers. The LAM/MPI 10.0 team will test these space science applications inside a software test environment for fault tolerance. The PI will provide advice to the LAM/MPI 10.0 team on application implementation issues and the software needs of space science image-analysis applications. The PI will attend the quarterly LAM/MPI 10.0 meetings. Expert optical and infrared astronomers will test the software, review the documentation and make suggestions for improvements. The project software products will be released to the general space-science and computing-science communities by placing them in the AISRP Code Archive Library. Project highlights highlights and annual reports will be posted at the main AISR website. Additional project related material will be posted on websites dedicated to this project on the web server of the National Optical Astronomy Observatory.

Automated Classification of X-ray Sources for Very Large Datasets

PI: Susan Hojnacki, Rochester Institute of Technology

The enormous amount of multidimensional data generated by increasingly higher quality observatories is resulting in very large astronomical databases. The growing data archives of the Chandra X-ray Observatory (CXO) and the X-ray Multi-Mirror Mission (XMM-Newton) provide excellent examples of of this conundrum. Unbiased source classification methods would augment and enhance standard spectral and temporal analysis techniques that are routinely applied to CXO and XMM data. We are proposing to develop a novel statistical source clustering and classification algorithm to maximize the science return from these and other large, multidimensional astronomical datasets. Our initial emphasis is on spectral classification of hundreds of X-ray sources detected in CXO observations of the Orion Nebula Cluster. We will generalize our classification techniques to take into account temporal attributes of X-ray sources and thereby create a robust clustering algorithm. The algorithm also will be extended for use in blindly classifying X-ray sources in other regions of the sky. Inputs and procedures specific to X-ray wavelengths will be modularized so the algorithm may be used with data from other regions of the electromagnetic spectrum. The ultimate goal is to develop the capability to group sources in large fields, such as a sky survey, independent of the requirement of a model or a priori knowledge of the nature of the sources (i.e., young stars, binaries, active galactic nuclei). Methods to visualize the results will be explored. The resulting algorithm will be developed into a tool and made available to the established centers for astronomical data analysis. The proposed research is relevant to Goal II, Astronomical Search for Origins, RFA 2(a). The final algorithm will increase science return from observational data through advanced knowledge discovery. The algorithm will be of immediate use to X-ray astronomers studying the mechanisms underlying X-ray emission to improve our understanding of of the nature and timescale of accretion onto young, solar-mass stars from protoplanetary disks.

Certificate based NVO wide Authentication and Authorization driven by provision of local database st

PI: Alexander Szalay, Johns Hopkins University

Recently the International Virtual Observatory Alliance (IVOA) has published a proposal for the basic building block of the Virtual Observatory, called an OpenSkyNode. This specifies a simple, queryable interface to the underlying archives of astrophysical data. This standard interface also enables interoperable services and portals built on top of the OpenSkyNodes. After studying the usage patterns of several large archives available at JHU (SDSS, SkyQuery), we have designed an experimental queue management system for large queries that also enables users to maintain their own value-added data next next to the main archives, and perform a step-wise analysis there. We have built an experimental Batch Query System with local user-owned databases (called MyDB). These ideas were presented at ADASS XIII in Strasbourg. We propose to further develop and extend our MyDB and batch query prototypes into multiple, distributed nodes of the Virtual Observatory. The introduction of asynchronous queries necessitates temporary user space as well as job submission and tracking facilities. Beyond building a reference application, we will also specify a standardized, generic SOAP interface for these tasks so that even if the engines for managing the jobs differ, a common interface would exist for submission and querying of jobs independent of platform and architecture. We would like to explore a distributed system whereby multiple nodes form a trusted network (Virtual Organization) and users with an account account (i.e. authentication) on one node would be accepted at other nodes without needing to register separately. We foresee users having datasets in multiple "spaces" on multiple machines. We propose to build a portal and appropriate protocols to enable the display of a given user's collections and the status of the respective workflows. This will require nodes to support queries about the user databases and jobs jobs held or submitted by users, and a certain level of authentication (this time of the portal) and authorization to make such queries. Interaction with other initiatives similar to MyDB, such as AstroGrids's MYSPACE will allow us to experiment with issues such as how exactly the Virtual Organization will work i.e. if I am authorized for MyDB by NVO should AstroGrid allow MYSPACE? In this area we shall collaborate with the AstroGrid community, a letter of support is included from Tony Linde project manager of AstroGrid. Savas Parastatidis of the University of Newcastle upon Tyne UK is particularly active in the field of security and web services and we are delighted to collaborate with him in this area. We are requesting funding for 3 years to cover approximately 3 person years of work. A personnel summary follows. All software generated through this research will be Open Source and fully available for anyone.

Stardust@home: A massively distributed public search for interstellar dust in the Stardust Interste

PI: Andrew Westphal, University of California Berkeley

In January 2006, the Stardust mission will return the first samples from a solid solar-system body beyond the Moon. Stardust was in the news in January 2004, when it encountered comet Wild2 and captured a sample of cometary dust. But Stardust carries an equally important payload: the first samples samples of contemporary interstellar dust ever collected. Stardust uses aerogel collectors to capture dust dust samples. Identification of interstellar dust impacts in the Stardust Interstellar Dust Collector probably cannot be automated, but will require the expertise of the human eye. However, the labor required for visual scanning of the entire collector would exceed the resources of any reasonably-sized research group. Here we propose to develop a project to recruit the public in the search for interstellar dust, based in part on the wildly popular SETI@home project, which has five million subscribers. We call the project Stardust@home.

Knowledge Discovery in a Virtual Universe

PI: Jeffrey Gardner, Carnegie Mellon University

Astrophysics is witnessing a flood of data from new ground and space based telescopes and surveys. With access to data sets spanning X-rays through to radio wavelengths we will soon be able to extract sources from any region of the sky and measure their properties across the full range of the electromagnetic spectrum. Individually, each of these data sets has the potential to advance our understanding of the processes that drive the formation and evolution of our Universe. It is, however, only when these data are combined that their full scientific potential will finally be realized; the scientific returns from the total will far exceed those from any one individual component. The recognition of this by the Astronomical community has led to a major new initiative: the National Virtual Observatory (NVO; http//www.us-vo.org). If we are to realize the full potential of the era of the NVO we must be able to explore, analyze and interact with these massive datasets. We need tools for knowledge discovery that not only make data delivery fast and easy to learn but also enable rapid data analysis on a massive scale. The resources required to analyze these terabyte and larger databases can be be found if we can harness the power of the upcoming generation of distributed computing resources (i.e. the Grid). The challenge we face is how do we help a user to realize the scientific value of the data without being limited by the laborious and complicated efforts required to implement data analysis techniques that span many processors distributed across a grid? Can we not only make data delivery fast fast and easy to learn (as the NVO is doing), but can we also minimize the time that it takes an astronomer to use that data to actually find an answer? This proposal will address the challenge of knowledge discovery in the era of the NVO and the Grid; providing the final component necessary to link the massive astrophysical data sources to massive distributed computational resources. We will develop new parametric and non-parametric algorithms and techniques that solve common problems in astrophysical data analysis and that are optimized for distributed computing environments. This research program has the potential to impact not only space science and the NVO, but also any NASA missions that generate large multidimensional datasets.

Precision High-Speed Crowded-Field Image Analysis in Space

PI: Matthew Lehner, University of Pennsylvania

We propose to develop a photometric reduction software package implementing the Expectation-Maximization (EM) Algorithm, an iterative Maximum Likelihood method. This algorithm is highly suited to PSF fitting on images of crowded fields where blending makes standard fitting routines difficult and time-consuming. Such a solution would be extremely important in cases such as analysis of crowded images with very large focal planes, or onboard analysis of images from space-based instruments where optimized useage of CPU cycles is paramount. The software will be capable of fitting galaxy shapes and a variety of PSF shapes.

Segmented Nonparametric Models of Distributed Data: From Photons to Galaxies

PI: Jeffrey D. Scargle, Ames Research Center

A novel technique, and an efficient algorithm to implement it, provides piecewise-constant models for a variety of one dimensional data types common in high energy astrophysics and cosmology, as well as in physical simulations of astrophysical systems. We will improve this algorithm in various ways (speed, parameter selection, alternative fitness functions, etc.) However, the major task will be extension of this this methodology to 2D data (such as galaxy surveys), 3D data (redshift surveys), and higher dimensional data spaces. We have recently discovered a way to transform higher dimensional problems into approximately equivalent 1D cases, solvable with the 1D algorithm. A major mathematical goal of this work is to find rigorous solutions in higher dimensions. We will simultaneously derive scientifically important results and hone the methods on analysis of GLAST photon maps, 3D galaxy positions derived from redshift surveys, and stellar infrared data useful for detection of molecular clouds. Ultimately these methods will be applicable to a wide range of astrophysical problems.

The Journey of the Sun-A Virtual reality Simulation: Date-Constrained Modelling and Visualization

PI: Priscilla C. Frisch, University of Chicago

The goal of this research is to construct effective computer grpahics tools to create a navigable 3D data-constrained model of the galactic environment of the Sun, and render the models using state-of-the-art computer visualization techniques. Sophisticated modern methods of eveidence assimilation, probabilistic reasoning, and inference will be combined with extant multi-spectral data to reconstruct unknown portions of the measured data, which can then be modified by the expert user.

Development of an IUE Time Series browser

PI: Derck Massa, SGT, Inc.

The International Ultraviolet Explorer (IUE) satellite operated successfully for 17 years. Its archive of more than 100,000 science exposures is widely acknowledged as an invaluable scientific resource that will not be duplicated in the foreseeable future. We have searched this archive for objects which were observed 10 or more times with the same wavelength coverage and spectral dispersion over the lifetime of the satellite. Using this definition of a time series, we demonstrate that roughly half of the more than 100,000 science exposures are members of such time series.identifying these datasets and developing a means to easily examine them in order to determine which objects varied over the lifetime of IUE would would be an extremely useful scientific asset. We have already identified the datasets and developed a prototype for a browser which operates over the web. The browser enables the researcher to visually inspect the repeated observations for variability and to examine each member spectrum individually. Further, once the researcher ascertains whether a specific dataset is worthy of further investigation, the entire dataset can be downloaded through links on the browser page for further scrutiny on the researcher's home computer. This project will produce a fully operational version of the browser.

Integrated Visual Simulation Pipeline - an extension of the Scientist's Expert Assistant

PI: Anuradha P. Koratkar, University of Maryland Baltimore County

We propose to research and prototype a fully integrated visual astronomical pipeline linking data archives, scientific analysis tools, and new proposal preparation systems. NGST funded the initial development of The Scientist's Expert Assistant (SEA) to research new visual approaches to proposal preparation. In the last call, AISR provided the seed funds to extend SEA with initial scientific modeling tools and generalizing the SEA model to multiple observatories. SEA's success to date has already led to its adoption by the Space Telescope Science Institute for production use with Hubble's observing program. SOFIA is in the initial stages of adapting SEA for their observing programs. SIRTF has adopted parts of the code and the MIDEX mission KRONOS is considering SEA as the baseline for for its user support strategy. Building on this experience, we plan to extend both the scientific simulation capabilities and the data archive visualization into an integrated pipeline. The goal is a simulation pipeline that will allow the user to manage the complex process of simulating and analyzing images without heroic programming effort. Tying this into SEA will allow astronomers to effectively come "full circle" from retrieving archival images, to data analysis, to proposing new observations. This simulation pipeline will increase scientific return within limited resources by improving the quality of observations and reducing the number of unusable observations. This is not only a valuable tool for traditional observatories and archives; it is also a key building block to allow the future Virtual Observatory to achieve its potential.

Novel Approaches to Supervised and Unsupervised Data Exploration

PI: David Bazell, Eureka Scientific

n this proposal we examine two novel approaches to the exploration and understanding of astronomical data: the use of unlabeled data for supervised classification and semi-supervised clustering. Because of the large and increasing volume of data from astronomical satellites and ground-based telescopes, researchers can no longer hope to examine the data by hand. Automated techniques are essential lest important discoveries be lost. Current automated classification methods rely on supervised learning algorithms, such as neural networks and decision tree inducers, that require training set containing large amounts of previously classified, or labeled data. While unlabeled data is often cheap and plentiful, using a human to classify the data is tedious, time consuming and expensive. We will develop methods whereby supervised classification techniques can make use of cheaply available, large volumes of unlabeled data to substantially improve their ability to classify objects. If the target classes are unknown, unknown, unsupervised clustering is a standard method of exploring unknown data and partitioning it into useful groups. We will also implement and explore several semi-supervised clustering methods. Finally, while classifier learning based on mixed labeled/unlabeled data and semi-supervised clustering can be viewed as separate problems, we will develop and evaluate a unified framework where the learned models directly provide clustering or classification solutions, or both, depending on the needs of the user. This approach allows assignment of astronomical data to predefined classes and will facilitate the discovery of new object classes. To demonstrate the utility of these methods, we will apply them to several test problems of current astronomical interest. Our primary scientific scenario will be aimed at identifying galaxy mergers in a variety of large (unlabeled) catalogs using labeled data from both simulations and observations. We will also examine morphological galaxy classification using data from from existing galaxy catalogs and the Hubble Medium Deep Survey. These several data sets contain different types of features used for classification. Both structural (e.g. shape and texture) and photometric features will be tested. The existence of different types of features is directly relevant to the co-training method, one of the methods we will investigate for building classifiers based on labeled and unlabeled data.

Parallel-Processing Astronomical Image Analysis Tools for HST and SIRTF

PI: Kenneth Mighell, National Optical Astronomy Observatories

This proposal describes a two-year research plan to develop and implement several parallel-processing astronomical image-analysis tools for stellar imaging data from the Hubble Space Telescope and the Space Infrared Telescope Facility. The new data analysis software tools will be written with the goal of enhancing the scientific return of the HST and SIRTF missions as well as future planned NASA astrophysical missions like the Next Generation Space Telescope. This project combines the enabling image-processing technology of the Principal Investigator's new digital PSF-fitting MATPHOT algorithm for accurate and precise CCD stellar photometry with enabling technology of Beowulf clusters which offer excellent cost/performance ratios for computational power. Data mining tools to do do quick-look stellar photometry and other scientific visualization tasks will also be written and used in order to investigate how such tools could be used at the data servers of NASA archival imaging data like like the Space Telescope Science Institute. Such data mining tools have the potential of making the serving of archival imaging data more efficient while simultaneously providing a more rewarding user experience for visiting astronomers. This project is structured into two one-year phases. The major software products (including original source code and detailed documentation) developed as part of this research project will be tested by expert optical and infrared astronomers before being released to the general astronomical community. The project software products will be posted (at the end of each one-year phase) on web sites dedicated to this project on the web server of the National Optical Astronomy Observatory.

The Journey of the Sun -- A Virtual Reality Simulatin: Data-Constrained Modeling and Visualization of Interstellar Matter Matter in our Galaxy

PI: Priscilla C. Frisch, University of Chicago

The richness of objects in the two-dimensional sky overhead belies the sparceness of space. The objective of this proposal is ti give context to this “empty” space by reconstruction and visualizing the Galactic surroundings of the Sun and nearby stars, including both spatial variations and temporal changes, such as stellar motions and the movement of interstellar matter (ISM) through space. The proposed research provides a new approach to data-interpretation by integrating multispectral ISM data (e.g., absorption and emission lines) with astrometric data on stellar motions to create coherent data-constrained models for 3D volumetric fields of the ISM in the Galactic neighborhood of the Sun. State-of-the-art computer techniques are used to manipulate, interact with, and visualize these models. The proposed tools and software can be extended to reconstruct 3-dimensional models of the volume distribution of any material giving rise to absorption lines, provided the distances of the background target objects are known. Our strategy is to merge cloud positions and velocities from the absorption data with two-dimensional emission line data (e.g., the HI 21-cm line) to fill in regions not sampled by the absorption lines. We will develop model visualization techniques that will use virtual reality methods augmented to handle continuous space-time navigation over scale sizes ranging from 1 AU to many megaparsecs. A single simulation could show the orbits of near-earth asteroids at a few AU from the Sun, the space-trajectories of the stars in Orion, the distribution of galaxies measured by the Sloan Digital Sky Survey and a 3D visualization of the Hubble Deep Field objects for which redshifts are available. The proposed research supports the National Virtual Observatory concept by providing computer-assisted methods for interpretation of ISM emission and absorption data, and by producing new methods for visualizing these data using virtual reality simulations and navigation tools adapted to the large spacetime scale ranges required for astronomical applications.

A High-Speed Data Access Component for a National Virtual Observatory Data Grid Node

PI: Ani Thakar, Johns Hopkins University

We propose to augment the Science Archive of the Sloan Digital Sky Survey and employ it as a testbed for the development of a high-speed data access layer for a National Virtual Observatory data grid node. Our goal is to produce a scalable, distributed archive framework that uses standardized software toolkits toolkits and interfaces so as to provide a prototype for grid-compatible astronomical archives. We aim to achieve this by developing a MPI-based parallel, distributed system that features loosely-coupled clusters of compute and data nodes. The current design of the SDSS Science Archive already features a high degree of parallelism and scalability. However, it uses a proprietary, non-portable communication protocol and data access primitives that are specific to the object-oriented DBMS that serves as its data repository. The proposed re-engineering aims to use MPI to provide high-speed communications and portability in order to obtain a data access layer that is optimized for parallel communication, independent of the underlying DBMS architecture, and conforms to the NVO standards and protocols. The virtual data grid that is expected to be the backbone of the NVO data services infrastructure will require individual archives to be structured in a data-parallel paradigm and be able to respond to requests requests for data and metadata in a fast, efficient, and standardized fashion. The successful deployment of the proposed component will also provide the high-speed data access required for the fast cross-identification and matching algorithms using co-operative agent technology that we are currently in the process of developing as part of an AISRP 2000 funded project. This proposal relates to the fabric fabric layer of the proposed NVO architecture, and requests funding for two years for partial salary support and travel support for the PI, travel support for one Co-I, and full support for one post-doc at JHU reporting to the PI. The project plan also includes collaboration with the Royal Observatory of Edinburgh (ROE) and the Edinburgh Parallel Computing Centre (EPCC) for MPI expertise. Additional travel support required for this collaboration is included in the budget.

Scalable Algorithms for Fast Analysis of Megapixel CMB Maps and Large Astronomical Databases

PI: Istvan Szapudi, University of Hawaii, Manoa

We propose to develop a suite of algorithms for spatial statistical analyses of future large astronomical data bases, such a megapixel CMB maps, galaxy catalogs, or any point source catalog. We will tackle hitherto unsolved computational challenges with high accuracy, such as: i) fast simulation of correlation functions and angular power spectra, ii) fast estimation of n-point correlation functions and iii) counts in cells statistics in from large CMB maps, angular, redshift and weak lensing surveys. We develop a unique interdisciplinary approach for algorithmic development, in which we reformulate the statistical problems arising in space astronomy with special attention to computational needs. This approach heavily relies on blending the principles of astronomy, statistics, computer and computational science, and balances statistical accuracy vs. practical computability.

Galaxy Photometry for NASA/IPAC Extragalactic Database

PI: James Schombert, University of Oregon

This project's goal is to provide a set of galaxy photometry tools to NASA/IPAC Extragalactic Database Database (NED). The system will allow new and experienced researchers to perform a full surface photometry analysis of any 2D data in NED's holdings or, by downloading the package, on their own optical or near-IR datasets. A heavy emphasis has been placed on interactive/visualization tools plus substantial documentation for the researcher new to the field of galaxy photometry. This project's modest budget will add NVO-type functionality to NED and act as a prototype to future expansions for data centers.

Classifying the High Energy Universe: A Prototype of the National National Virtual Observatory

PI: Thomas A. McGlynn, Goddard/USRA

We propose to prototype the National Virtual Observatory, building the tools and protocols needed to integrate large, distributed datasets to do science infeasible with a single institution's resources. Our pilot pilot research project is to build an automated classifier for X-ray sources and to use it to try to distinguish the physical classes of all known X-ray objects. This project is carefully selected as challenging but feasible, excellent science in its own right, and exercising the key technologies for the NVO. Massive datasets from the HEASARC, ST-MAST, and Chandra archives along with information information from VizieR, 2MASS, FIRST and other systems will all be needed within this effort. The deliverables will include systems for coherent access and integration of information from many different astronomy data providers, the methodology and software for the creation of automated classifiers in the NVO regime, a realistic assessment of the current connectivity of astronomical sites, and our X-ray classifier itself. Our clear science context will help ensure that the tools we develop are truly useful to the astronomy community. The science goals are particularly timely given the recent launches of the Chandra and XMM-Newton observatories. The number of known X-ray sources will soon double, but our current tools for categorizing them are extremely tedious and resource intensive. Our classifier must handle quite heterogeneous data. We propose using a network of sub-classifiers of several different types to address the diversity of data available. This proposal leverages substantial institutional commitments from the collaborating groups. Funding is requested only to support the actual actual software development activities involved in this project. The science oversight and astronomical research activities are provided by the collaboration from other institutional resources.

Cosmic Microwave Background Analysis Tools

PI: Julian Borrill, Lawrence Berkeley National Laboratory

The Cosmic Microwave Background radiation provides a unique picture of the early universe. To realize realize its scientific potential NASA is leading an international effort to obtain precise measurements of the CMB temperature and polarization from ground-based, balloon-borne, and satellite observatories. This proposal supports the development and distribution of the novel computational algorithms and implementations needed to plan observing strategies, simulate observations, and analyze data. The CMB temperature power spectrum shape is a strong constraint on cosmologies. Its measurement requires a large area of sky observed at high resolution by many detectors at many frequencies. The volume of such data makes their analysis a serious computational challenge; our MADCAP software can can analyze maps with O(100,000) pixels but new data will give up to O(10,000,000) pixels, requiring new approximate algorithms. New methods are also needed to analyze multiple channels to account for systematic effects from their cross-correlation and their different frequencies and beam sizes. CMB polarization provides a new window into the very early universe. Currently in the detection phase, the challenge is to observe a signal much fainter than the temperature anisotropies, requiring well-planned observations and novel analyses. We will develop polarization extensions to our FORECAST and WOMBAT flight-planning tools, as well as first-generation simulation and analysis tools. To compare the CMB temperature or polarization power spectrum with cosmological models we search a 10-20 dimensional space of correlated parameters to get both the maximum likelihood parameter combination and the full likelihood contours. Since the parameter-likelihoods are non-Gaussian formally we need to calculate them throughout this infinite space. In practice we must restrict ourselves to a finite sampling of a finite subset of the space: our goal is to codify and optimize the appoximations in this restriction.

Bibliography

Making maps of the cosmic microwave background: The MAXIMA example Stompor et al.; Physical Review, (2002) D 65, 0022003

An Automated System for the Reduction and Analysis of X-ray Data from Galaxies

PI: Andrew F. Ptak, Johns Hopkins University

Much an observer's time is spent on repetitive data reduction, and likewise certain aspects of data analysis such as determining errors for spectral fits is also repetitive. Often, particularly early on in a mission's life, data reduction and analysis need to be repeated as calibration data and tools are updated. We propose the development of a system for the automation of the reduction and analysis of X-ray data from galaxies for the missions ASCA, ROSAT, AXAF and XMM. For this specific scope, the system will handle all of the tasks that could appropriately be handled in an automated fashion. This will free a considerable part of the astronomer's time for more advanced analysis of the data products. Furthermore Furthermore it is likely that it will not be feasible for astronomers to complete many interesting projects without such a system in place given the large amount of time required and the volume of the data involved in X-ray research. This problem will become increasingly acute as future X-ray missions produce more voluminous and complex data sets. The system will be based mostly on scripts that drive existing software (such as FTOOLS and XSPEC) but will also consist of advanced (also scriptable) software already in development at CMU for spatial and temporal analysis of complex X-ray data. A large emphasis will be placed on ease-of-use and extensibility. This system will be useful for both "legacy" analysis of a given galaxy (i.e., a complete spectral, spatial and temporal analysis of the publically available data from the supported missions) and survey studies (analysis of deep fields and searches for serendipitous sources). A primitive early version of the system has already been successfully used to analyze and re-analyze the ASCA data from a sample of galaxies. As the availability of funding for researchers is limited, a system such as this will be of considerable value in assuring that as much X-ray data as possible in the archives and forthcoming from missions such as AXAF and XMM are analyzed and published. Following completion of the initial version of the system, the system will be expanded, e.g., to include other extragalactic data such as clusters and groups, groups, and to include tools for multi-wavelength data analysis. The successful completion of these goals will be particularly useful to institutions will a limited number of X-ray-proficient personnel who nevertheless wish to take advantage of the X-ray data available from these (and future) missions. Since the products of the system will include "calibrated" images, spectra and lightcurves as well as "physical" "physical" parameters from fits to these products, such as temperatures and spatial extents, the system could be applied to entire archives to produce databases for use with future data-mining projects. We will also explore the visualization of the final results (particularly spatial results), to aid in the scientific understanding of the data and for public outreach.

Advanced Galaxy Cluster Detection Software

PI: Robert Nichol, Carnegie Mellon University

The Sloan Digital Sky Survey (SDSS) is one of the first of many NASA sponsored missions to produce a trillion pixel image of the sky. A main objective of the SDSS is to determine the spatial clustering of galaxies in the universe. With the advent of large, high-precision surveys like the SDSS, it is now time to apply advanced algorithms taken from statistics, signal processing and computer science to these data to quantify the degree of clustering seen and thus robustly find clusters of galaxies within these large datasets. Over the past five years, the authors of this proposal have collaborated with statisticians, engineers and computer scientists to test a variety of new algorithms on astronomical data. We now come together to produce -- for the benefit of the community -- a suite of high performance cluster detection software that synthesizes the best of all our experiences. In addition, we will use this suite of software to produce the definitive galaxy clustering database for the SDSS. This will be achieved as follows: First, we will develop the test bed for our detection code. This will consist of publicly available software for generating simulated data sets with known clustering properties. These simulations are critical for defining the selection function of any cluster survey. Second, we will compare the different detection algorithms that have been developed by the team members. Based on these comparisons, we will select the best combination of detection algorithms, and implementation techniques, and use these to produce an open source software suite. Finally, we will apply the software we have developed to the SDSS galaxy database and produce a robust, well-understood and well-parameterized catalog of many thousands of clusters. The broad applicability of this software is critical to its success and should be useful for the Large Synoptic Survey Telescope, the "Virtual Observatory" and many other NASA missions like 2MASS, GALEX, Chandra.

Bibliography

THE CUT-AND-ENHANCE METHOD: SELECTING CLUSTERS OF GALAXIES FROM THE SLOAN DIGITAL SKY SURVEY COMMISSIONING DATA Goto et al.; AJ, (2001) 123 pp.

DETECTING CLUSTERS OF GALAXIES IN THE SLOAN DIGITAL SKY SURVEY. I. MONTE MONTE CARLO COMPARISON OF CLUSTER DETECTION ALGORITHMS Kim et al.; AJ, (2001) 123 pp. 20

WITS the WEB Infrared Tool Shed

PI: Mark G. Wolfire, University of Maryland College Park

Using a previous NASA ADP type 2 grant, we created a well known and user tested WEB based tool for for the analysis of dust continuum and PhotoDissociation Region (PDR) observations. The WEB Infrared Tool Shed (WITS) contains two "toolboxes", the PDR Toolbox (PDRT) and the Dust InfraRed Toolbox (DIRT) and are currently available at http://dustem.astro.umd.edu or http://wits.ipac.caltech.edu. These toolboxes provide an extensive data base of PDR and dust continuum models which can be "mined", displayed, and manipulated using a Java user interface. This proposal seeks to enhance these tools with an expanded data base of model sources, with additional algorithms and user options to find the best fit models for an input data set, and with modules tailored for the spectral and spatial responses of specific NASA instruments.

Multifrequency, Multiresolution Image Detection

PI: Andrew John Connolly, University of Pittsburgh

The next decade will mark the start of a new era in multi-frequency astronomy. New surveys and satellites are coming on line that will open up the full electromagnetic spectra at higher precision and to fainter flux levels than ever before (from X-rays through to the submillimeter and radio). Together these multi-frequency surveys will provide a bolometric view of the Universe from which we might fully reconstruct the properties of Galactic and extra-galactic sources. Studying these data sets in isolation will provide some insight into the processes that drive the evolution of our Universe. It is only, however, however, if we analyze the multi-frequency data in unison that we might fully exploit their scientific potential. We have, therefore, developed a new and novel technique for combining multi-frequency data that provides objective source detection. We propose here to develop these algorithms into a suite of software tools that will enable the astronomical community to explore multi-frequency, multi-resolution data in an optimal fashion.

Bibliography

The Angular Correlation Function of Galaxies from Early Sloan Digital Sky Survey Data
Connolly et al.; ApJ, (2002) 579 pp. 42

A Parallel Genetic-Algorithm-based approach to White Dwarf Pulsation Models

PI: Donald E. Winget, University of Texas Austin

White dwarf asteroseismology offers the opportunity to probe the structure and composition of stellar objects governed by relatively simple principles. The observational requirements of asteroseismology have been addressed by the development of the Whole Earth Telescope (WET) in conjunction with the HST, but the analytical procedures need to be refined before this technique can yield the complete physical insight that the data can provide. We propose to develop an optimization method utilizing a genetic algorithm (GA) for fitting white dwarf pulsation models to the observations. We will develop software to use this global optimization approach in two distinct ways: (1) to investigate the uniqueness and objectivity of the solutions by combining existing models with a GA-based fitting routine, and (2) to investigate the completeness and adequacy of our understanding of the principles governing white dwarf interiors by parameterizing the constitutive physics and using a GA to find the family of solutions that produce observationally indistinguishable behavior. We have configured a specialized computational instrument to develop this software---a parallel metacomputer consisting of a network of of 64 minimal PCs running Linux. The structure of a GA is very conducive to parallelization, so this metacomputer will allow us to develop and run the code on a practical timescale.

Bibliography

12C(alpha, gamma)16O from WD Asteroseismology Metcalfe; Salaris; Winget; ApJ, (2002) 573 pp. 803